Rev.Phil.Psych.
DOI 10.1007/s13164-009-0016-1

Survey-Driven Romanticism

Simon Cullen

Abstract Despite well-established results in survey methodology, many experimental philosophers have not asked whether and in what way conclusions about folk intuitions follow from people's responses to their surveys. Rather, they appear to have proceeded on the assumption that intuitions can be simply read off from survey responses. Survey research, however, is fraught with difficulties. I review some of the relevant literature—particularly focusing on the conversational pragmatic aspects of survey research—and consider its application to common experimental philosophy surveys. I argue for two claims. First, that experimental philosophers' survey methodology leaves the facts about folk intuitions massively underdetermined; and second, that what has been regarded as evidence for the instability of philosophical intuitions is, at least in some cases, better accounted for in terms of subjects' reactions to subtle pragmatic cues contained in the surveys.

1 Experimental Philosophers' Methodology

We will be concerned with a methodology that is common to a great deal of work that has gone under the label of experimental philosophy. To be sure, not all experimental philosophers employ or endorse this methodology; but its use is widespread, so carefully examining it is an important task.

S. Cullen (✉)
Princeton University, Princeton, NJ, USA
e-mail: scullen@princeton.edu

Published online: 30 January 2010

How do experimental philosophers describe what they are doing? Here is what Nadelhoffer and Nahmias say:

> Experimental [philosophers] use the *methods of experimental psychology* to probe the way people make judgments that bear on debates in philosophy (2007, p. 123, my emphasis).

Here is Shaun Nichols:

> Recently, researchers have begun to exploit *social scientific methodologies* to characterize folk concepts. … A second approach … applies the *methods of cross-cultural psychology* to philosophical intuitions (2004, p. 154, my emphasis).

And Nahmias et al.:

> Philosophers working in the nascent field of 'experimental philosophy' have begun using *methods borrowed from psychology* to collect data about folk intuitions … (2005, p. 123, my emphasis).

And Joshua Knobe:

> The new field of experimental philosophy seeks to subject [philosophers' claims about folk intuitions] to rigorous tests using the *traditional methods of cognitive science*—systematic experimentation and statistical analysis (2007, p. 81, my emphasis).

And Alexander and Weinberg:

> Experimental philosophers are unified behind … the application of *methods of experimental psychology* to the study of the nature of intuitions (2007, p. 56, my emphasis).

There are many more examples, but the reader will already have noticed the common theme: experimental philosophers say they're doing *science*. They are applying methods "borrowed from psychology"—or "social science", or "cognitive science", or "cross-cultural psychology", or "experimental psychology", or … —to the study of folk intuitions. It's time we asked, Is this really so?

In their seminal contribution to the "restrictionist programme" (Alexander and Weinberg 2007), Weinberg et al. (2006) characterised philosophers who rely on their own intuitions to support normative philosophical claims as "Intuition Driven Romanticists". Intuition Driven Romanticists in the theory of knowledge hold that

> knowledge of the correct epistemic norms … is implanted within us in some way, and with the proper process of self-exploration we can discover them (ibid., p. 194).

Weinberg et al. are concerned that analytic philosophers continue to naïvely assume—indeed, in the face of apparently clear empirical data to the contrary—that they can read off facts about what 'we' find intuitive simply by consulting their own intuitions. While I share their concern, I believe there is a

parallel problem with experimental philosophers' methodology—which I call "Survey-Driven Romanticism". According to Survey-Driven Romanticism,

> people's philosophical intuitions are implanted within them in some way, and by administering simple surveys we can discover them.

Examples where experimental philosophers appear to directly read off folk intuitions from survey responses abound, so my choice must be somewhat arbitrary.

Nahmias et al. take it that

> negative responses [to a survey depicting an agent in a deterministic world] indicate the intuition that determinism conflicts with free will and responsibility (2005, p. 565).

Joshua Knobe claims to

> begin with some *straightforward data* about people's intuitions concerning specific cases (2006, p. 205, my emphasis).

Swain et al. claim to have *demonstrated* that

> intuitions vary according to [the order in which] thought experiments are considered (2008, p. 153).

Weinberg et al. claim that

> what we've been reporting are just the *brute facts* that intuitions in different groups differ (2006, p. 215, my emphasis).

Machery et al. simply assert that their surveys

> modeled on Kripke's Gödel case … elicit culturally variable intuitions (2004, B7).

Machery et al. consider the

> extreme but very live possibility [that] individuals might have [certain] intuitions on some occasions and [conflicting intuitions] on other occasions [ibid., B8],

but they hardly pause for the more mild possibility that the differences in their subjects' survey responses mightn't reflect underlying differences in philosophically-interesting intuitions at all.

All this has been a serious mistake; we must distinguish intuitions from survey responses. This does not require any detailed account of what intuitions are, so let us grant that they are judgments whose contents are "singular classificational propositions, to the effect that [X is a case of C]" (Goldman and Pust 1998, p. 182). Survey responses are a kind of behaviour generated by several different inputs, only one of which might be a judgment of the kind described by Goldman and Pust. Other inputs are the background beliefs a respondent draws on when interpreting the vignette; her beliefs about what

the researchers are interested in; her sensitivity to conversational norms; and so forth.

Experimental philosophers can effectively equate survey responses with intuitions only by ignoring the established social and cognitive science literature on survey methodology. In my review of the growing body of experimental philosophy literature I have found just one reference to a serious discussion of survey methodology.[1] The conclusion to be drawn is that, despite their pretensions, many experimental philosophers have given no serious thought to methodology. This not only undermines their claim to be doing science, as we shall see, it often leaves the philosophical significance of their findings unclear.

It is true that when a philosopher says "We find X intuitive" ("we find it natural to say X"), she does not mean "I find X intuitive", and she does not, in general, mean "We philosophers find X intuitive". Rather, she means that competent, thoughtful speakers would, under certain circumstances, find X intuitive. Experimental philosophers are right to emphasise that appeals to intuition typically engender empirical commitments. And they are right to suspect that what philosophers find intuitive might diverge, possibly quite often, from what lay-people find intuitive. Our concern, then, is not with the aims of experimental philosophy; rather, it is with experimental philosophers' methods and the deeply mistaken assumptions which underlie them.

2 Asking Different Questions Yields Different Answers

Antti Kauppinen recently drew attention to the dubiousness of the unmediated inference from survey results to intuitions. In his (2007) critique of experimental philosophy, Kauppinen hypothesised that both pragmatic and semantic factors play an important role in generating subjects' responses to experimental philosophy surveys. If experimental philosophers are interested in folk concepts, he argued, then they need to distinguish the genuinely semantic contribution to survey responses from the influences of pragmatic distractions.

On this score I agree with Kauppinen, and the experiments presented here support his concern. But he is pessimistic that semantic and pragmatic factors can be experimentally distinguished:

> On Grice's view, conversational implicatures are essentially such that they can be worked out, given the assumption of conversational cooperation and facts about the context, including mutual beliefs … But

[1] Searching all of the electronic scholarly resources accessible from Melbourne University—which include all major philosophy, psychology, social and cognitive science journals—for the terms ["experimental philosophy" "survey methodology"] turned up only two results, neither of which contained a discussion of survey methodology. Searching for [Schwarz "Experimental philosophy"] yielded two additional results: Goldman and Pust (1998) who cite Schwarz (1995) in their *defence* of intuitions; and Doris et al. (2007) who cite Schwarz (1996).

> when a person responds to a yes/no survey question (or rates assent on a Likert scale), just what is the conversational context? Who is he or she conversing with, and how do we work out what he or she assumes about the hearer's beliefs? Frankly, this is a baffling task. Once again, an actual dialogue would help, but it would mean leaving behind the Survey Model and its pretension to scientific objectivity (ibid., p. 107).

Agreed: Questions about how the conversational context of the experimental setting and the formal features of the survey influence subject responses urgently need to be worked out. But I'm not sure why Kauppinen thinks they are so intractable.

Here is one answer to Kauppinen's questions which social and cognitive scientists have developed in considerable detail over the past two decades: the context in which a person answers a yes/no survey question or rates assent on a Likert scale is essentially that of *a conversation*—albeit a rather one-sided one—between researcher and experimental subject. Such conversations proceed according to the usual conventions of language use and the ordinary norms of conversation (Schwarz 1995, 1996). It is on this self-consciously Gricean model which survey researchers have developed the "science of asking questions" (Schaeffer and Presser 2003).[2]

Norbert Schwarz provides a beautiful demonstration of how non-semantic features can profoundly impact how subjects interpret a survey questionnaire. Consider the following two formally equivalent rating scales:

0	1	2	3	4	5	6	7	8	9	10
−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5

Schwarz's surveys asked subjects "How successful would you say you have been in life?" In one survey the response scale ranged from 0 = 'Not at all successful' to 10 = 'Extremely successful'; and in a second survey the scale ranged from −5 = 'Not at all successful' to +5 = 'Extremely successful'. Schwarz found that "whereas 34% of the respondents endorsed a value below the mid-point of the 0 to 10 scale, only 13% endorsed one of the formally equivalent values on the −5 to 5 scale" ($p < .001$) (Schwarz 1995, p. 157). The phrase "not at all successful" is ambiguous between an absence of noteworthy successes and the presence of explicit failures. Licensed by conversational norms, respondents turn to the context of the question—the numeric labelling—to infer which meaning the researcher has in mind.

It is crucial to note that without an appreciation of how respondents draw on contextual information to determine the meaning of the question (viz., the researcher's speaker meaning), these results would appear to indicate that responses are highly sensitive to "irrelevant factors". But they are not. The apparently straightforward question "How successful would you say you have

[2]For an overview, see Krosnick (1999).

been in life?" asks something quite different in different contexts. It is not literal meaning which matters to respondents—it's speaker meaning, and this crucially depends on context.

* * *

Experimental philosophy surveys provide many opportunities for unwittingly communicating valuable contextual information to subjects. Consider, for example, the following survey vignette which Weinberg et al. used in their (2006) study on epistemic intuitions:[3]

> Dave likes to play a game with flipping a coin. He sometimes gets a "special feeling" that the next flip will come out heads. When he gets this "special feeling", he is right about half the time, and wrong about half the time. Just before the next flip, Dave gets that "special feeling", and the feeling leads him to believe that the coin will land heads. He flips the coin, and it does land heads.
>
> Did Dave really know that the coin was going to land heads, or did he only believe it?
>
> REALLY KNOWS ONLY BELIEVES

Reading this survey, you may have noticed a number of peculiarities. First, the words *special feeling* are consistently enclosed in quotation marks; and since these are used according to several different conventions which differ in their implicatures, this introduces a certain ambiguity. Do Weinberg et al. mean to quote Dave directly? Or are they expressing in print what we might express in speech with exaggerated articulation or by adding the words 'so-called' in the appropriate places? To people who find the latter interpretation most natural, Weinberg et al. would seem to be very purposefully expressing their own dim opinion of the epistemic propriety of Dave's special feelings.

Second, the intensifying adverb *really* in the question ("Did Dave *really* know …") and in the first response alternative ("REALLY KNOWS") raises the standard for answering that Dave knew the coin would turn up heads. Conversely, the downtoning adverb *only* in the question ("… or did he *only* believe it?") and in the second possible answer ("ONLY BELIEVES") lowers the standard for answering that Dave only believed.

Third, presenting "only believes" as an alternative response to "really knows" communicates Weinberg et al.'s own belief that the distinction between *real* knowledge and (mere) belief is highly pertinent.

Though little differences like these might seem trivial to you and me, to someone struggling to understand and intelligently respond to an experimental philosophy survey they could be richly meaningful. Subjects rely on the assumption that we researchers are cooperative communicators, that we mean to provide for their consideration only what we believe to be relevant to their

[3]Nichols et al. use the same survey in their (2003), as do Swain et al. (2008).

task of generating responses. This leads subjects to look for meaning in the survey's form as well as its explicit content.

The experiments presented in the next section all concern examples where the survey's formal structure systematically influences subjects' responses, often strikingly. But before we turn to them, I want to suggest a plausible explanation for the power pragmatic influences wield on experimental philosophers' survey results. Many experimental philosophers want their "intuition probes—the cases that we ask subjects to judge—to be similar to cases that have actually been used in the recent literature" (Weinberg et al. 2006, p. 202). Thus, following in the tradition of contemporary thought-experiments, experimental philosophers have invented some stunningly bizarre survey vignettes. Such vignettes are probably harmless within the explicit conversational contexts in which philosophers consider thought-experiments, and they make doing armchair philosophy fun. But when transposed into the far more ambiguous context of an experimental survey, continuing this tradition has been a disaster. Convoluted thought-experiments are ripe with opportunities for confusing subjects, breaching conversational norms, and inadvertently conveying information which, in their effort to provide intelligent responses, subjects mistakenly interpret meaningfully.

The problem is doubly serious for experimental philosophers since people tend to rely most heavily on contextual and pragmatic cues when the meaning of the survey or their task is unclear (see, e.g., Schwarz 1995, 1996; Strack et al. 1988; Krosnick 1999). Indeed, experimental philosophy subjects are *ipso facto* at a significant disadvantage since it is often a precondition of their participation that they have no idea why anyone would be interested in finding out what the folk think about Gettier scenarios, much less what a Gettier-scenario actually is (see, e.g., Machery et al. 2004, B8-9).

2.1 Are Intuitions Sensitive to the Order in Which Cases Are Considered?

In the previous example from Weinberg et al.'s study, I noted several features, internal to both the vignette and question, which subjects might draw on in an effort to understand and respond to an experimental philosophy survey. Subjects can also extract meaning from the context *surrounding* the question, which can include other vignettes included in the survey. In their recent paper, Swain et al. (2008) exploit this sensitivity to argue for another "irrelevant factor" to which intuitions are supposedly sensitive: intuitions vary, they maintain, according to the order in which cases are considered.

In Swain et al.'s experiment one group of subjects were primed with a vignette about Karen, "a distinguished professor of chemistry". Karen learns from "a leading scientific journal that mixing two common floor disinfectants, Cleano Plus and Washaway, will create a poisonous gas that is deadly to humans" (p. 154). As a result, Karen comes to believe that mixing these disinfectants creates a poisonous gas; and the idea is, her belief is an obvious case of knowledge. A second group of subjects were primed using Weinberg et al.'s story about Dave, who occasionally gets a special feeling as

to whether a coin will turn up heads or tails. The idea is, Dave's belief that the coin will next turn up heads is an obvious case of non-knowledge. Subjects in both groups were asked to indicate on a 5-point Likert scale the extent to which they agreed or disagreed with a claim attributing knowledge to either Karen or Dave.

Having been primed with one of the vignettes and the task of assessing the epistemic status of its protagonist's belief, subjects in each group responded to a more problematic vignette based on Lehrer's (1990) Truetemp thought-experiment:

> One day Charles was knocked out by a falling rock; as a result his brain was "rewired" so that he is always right whenever he estimates the temperature where he is. Charles is unaware that his brain has been altered in this way. A few weeks later, this brain rewiring leads him to believe that it is 71 degrees in his room. Apart from his estimation, he has no other reasons to think that it is 71 degrees. In fact, it is 71 degrees.
>
> Please indicate to what extent you agree or disagree with the following claim: "Charles knows that it is 71 degrees in his room".
>
> ☐ Strongly agree, ☐ Agree, ☐ Neutral, ☐ Disagree, ☐ Strongly disagree

Swain et al. found that reading about Dave's special feelings first means you'll be a little more likely to accord Charles knowledge than had you not read about Dave, and reading about Karen first means you'll be a little less likely to accord Charles knowledge than had you not read about Karen. Swain et al. conclude: subjects' "intuitions *about the Truetemp case* reverse direction depending on whether the case is presented after a case of clear non-knowledge" (ibid., p. 144, my emphasis).

How do Swain et al. arrive at this conclusion? By manipulating certain features of their experiments which they deem "irrelevant", they are able to affect a slight but statistically significant influence on the mean response to Charles' case. On this basis they conclude that their subjects' "intuitions track more than just the philosophically-relevant content of the thought-experiments". Now it is surely true, as Swain et al. have demonstrated, that *survey responses* track more than just the philosophically-relevant content of the vignettes—but it is a fallacy to conclude on this basis alone that *intuitions* do too.

Schwarz et al. (1991) and other survey researchers have developed a sophisticated framework for understanding assimilation and contrast effects in part-whole question sequences in terms of Gricean conversational norms. A part-whole question is one where a specific question is followed by a general one. For example, if I ask "Do you enjoy eating junk-food?" immediately after having asked whether you enjoy eating Jelly-beans, you will interpret my second question as asking "Do you enjoy eating junk-food *other than* Jelly-beans?" The reason for this appears to be rather simple: if you interpret the second question literally it constitutes a request for redundant information

(since Jelly-beans are a subset of junk-food). This would violate Grice's Maxim of Quantity: speakers are assumed to make their requests and contributions to conversation as informative as required and not more so. Having been primed with the specific question, you simply interpret the general question in line with everyday conversational norms—as requesting new information. The implied meaning of the question "Do you enjoy eating junk-food?" varies depending on which question precedes it.

I want to suggest that something similar is responsible for Swain et al.'s results: the *implied meaning* of the primed question about Charles varies depending on the context in which it occurs. It is not that the subjects in each group have different intuitions *about Charles' case*, they effectively respond to different questions. On this account, Swain et al.'s subjects tried to provide meaningful, informative responses to the survey questions. They assumed, quite reasonably, that the researchers know that Dave's is not a case of knowledge, and that the researchers know that the subjects know this (*mutatis mutandis* for Karen). Given that the researchers aim to make all of their contributions to the conversation relevant as Grice's co-operative principle requires, that they have asked such a *prima facie* obvious question needs explaining. Happily for subjects, an explanation becomes apparent as soon as they encounter Charles' more problematic case: *the researchers want me to compare Charles' case to Dave's*. (Intelligent subjects could hardly fail to notice that the cases are purposefully chosen to contrast with one another!) On this reading, Swain et al.'s subjects answer that Charles' case is more obviously a case of knowledge than Dave's and less obviously a case of knowledge than Karen's. And of course they're right.

Swain et al. manipulate the context in which their subjects judge the Truetemp case so as to present those judgments out of context and conclude that they are influenced by "irrelevant factors"—"fairly minor and recent perturbations in their cognitive environment" (Alexander and Weinberg 2007, p. 67). It may well be true that the order in which a series of cases is considered should not influence the applicability of a concept to any one of those cases. Swain et al.'s results are consistent with this. What Swain et al. have not considered is that within the conversational context in which their subjects consider Charles' case, the ordering is highly relevant: it helps to determine the very meaning of their subjects' task.

I think this account is *prima facie* plausible—but we are doing experimental philosophy. So how can we test it? Strack et al. (1988) demonstrated how Grice's Maxim of Quantity mediates question order effects in part-whole question sequences. They were able to effectively switch the non-redundancy norm off by manipulating the conversational context with explicit lead-in questions. Though Swain et al.'s surveys are more subtle, on the reading I am urging their results are also the effect of conversational norms guiding subjects' interpretation of the task. Thus it should be possible to substantially reduce the contrast effect by manipulating the conversational context in which subjects consider the cases.

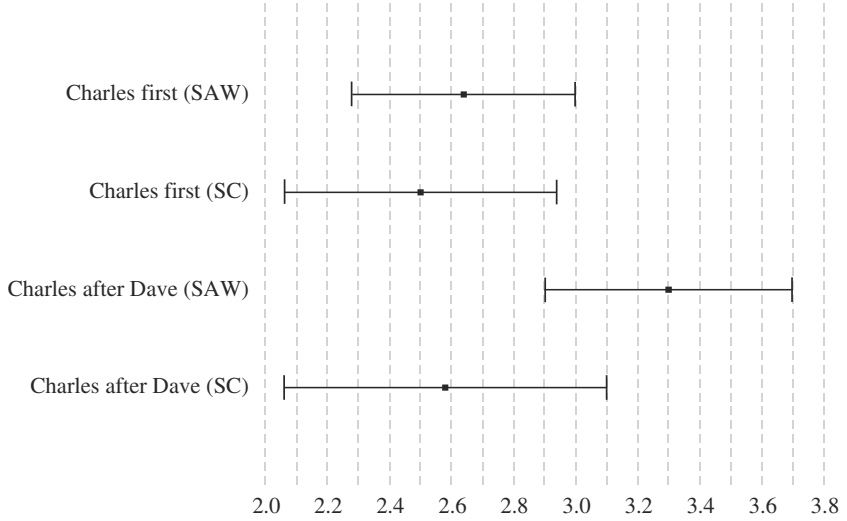

Chart 1 Mean response to Charles: Swain et al. and SC showing 95% CIs. CIs for Swain et al.'s means were estimated from their charts

In Swain et al.'s experiment the instructions to subjects were:

> We are investigating what different people's opinions are about knowledge. In each question, please indicate to what extent you agree or disagree with that statement.

To test the present hypothesis I prefaced Swain et al.'s original Dave-then-Charles survey with the following lead-in:[4]

> We are designing a survey for a cognitive science experiment exploring people's concept of knowledge. This involves testing prospective questions.
>
> We are interested in your response to two questions, one of which might be included in the final survey. Please consider each independently.

When this lead-in is used, the average response to Charles' case after Dave's is 2.58.[5] This is not significantly different to the baseline Swain et al. established

[4]By far the most significant effect Swain et al. uncovered is Dave's case on Charles' ($p = 0.043$). It therefore makes sense to test this case first.

[5]Responses are coded from 1 (strongly disagree) to 5 (strongly agree). Prospective subjects were approached at two Melbourne cafés. The experimenters introduced themselves and read the lead-in which was also printed at the top of the page. No demographic data was collected.

when they tested subjects' reactions to Charles' case considered first, and which I have closely replicated (see Chart 1, Charles first (SC)). Comparing subjects who considered Dave's case before considering Charles' (N = 26; mean response to Charles = 2.58 [95%CI: 2.06–3.01]) to those who considered Charles' first (N = 24; mean response = 2.5 [95%CI: 2.94–2.06]) shows that subjects' judgments about Charles' case are not affected by first considering Dave's ($p = 0.8$). Further experiments are clearly needed (and underway), but it appears from this preliminary data that when the conversational context is made explicit in this way, the priming effect disappears.

The present hypothesis also explains why Swain et al. found that responses to Charles' case did not shift when subjects considered it after responding to a more problematic case "expected to generate mixed intuitions with some subjects willing to attribute knowledge, and others not" (2008, p. 143). Because this problematic case is less extreme than either Karen's or Dave's, the contrast between it and Charles' case is also less extreme. Thus it is not clear to subjects what, if any, the comparison is supposed to be. The present hypothesis predicts that under these conditions there will be no substantial shift in responses to Charles' case (cf. Schwarz et al. 1991, p. 19), which is exactly what Swain et al. found.

Our purpose, however, is not to provide a definitive pragmatic account of Swain et al.'s data, but to consider the methodology and assumptions which underlie it. Swain et al. draw the following meta-methodological conclusion from their results:

> The fact that people's intuitions about particular thought-experiments vary based on what other things they have been thinking about recently is troubling. Philosophers who rely on thought-experiments should be especially concerned about findings that indicate that, at least in some cases, subjects' intuitions are easily influenced (ibid.).

Swain et al. (and many others) clearly assume that the context in which their philosophically innocent and experimentally naïve subjects consider the Truetemp case is in *all relevant* respects identical to that in which philosophers consider Lehrer's Truetemp thought-experiment. Here, then, are just three crucial differences which Swain et al.'s methodology must address:

1. Swain et al.'s subjects must try to understand why on earth anyone would ask such a bizarre, hardly intelligible question as whether a unreflective human thermometer knows that it's 71 degrees. Philosophers understand that the only salient fact about Charles is that he has no introspective access to the basis of his beliefs.
2. Swain et al.'s subjects try hard to provide intelligent responses to the surveys, taking into account the likely interests of the researchers (see, e.g.,

Norenzayan and Schwarz 1999; Schwarz 1995, 1996).[6] Philosophers simply try to respond to the thought-experiment as they understand it.

3. Swain et al.'s subjects interpret the survey vignette in a highly ambiguous conversational context. Philosophers understand the point and purpose of considering thought-experiments—and as we have seen, priming effects can disappear when the conversational context is made explicit (cf. Strack et al. 1988; Schwarz et al. 1991).

Swain et al. therefore provide little reason to believe their "data impugning various intuitions to present a real challenge for philosophers who wish to rely on intuitions as evidence". Rather, they provide a clear example of the fallacy of Survey-Driven Romanticism, that one may read off subjects' intuitions about the philosophically interesting features of unusual, unexplicated scenarios, from their survey responses.

3 Survey Pragmatics: Response Alternatives

We have so far examined how changes to the context surrounding a survey can affect how subjects interpret its vignette and questions. The following experiments examine how the formal structure of the survey itself can bear on subjects' responses.

3.1 Methods

Predominantly high-school aged (72% [95% CI: 69 – 75]) North American (93% [95% CI: 89 – 97]) subjects participated in surveys using the online survey website "Quibblo". Subjects selected surveys by searching for topics and keywords of interest or by browsing the titles of recently posted or popular surveys.

Concerns about possible biases meant that experiments were replicated throughout the study. The results of these replications were highly consistent (see Chart 2). The online appendix contains a discussion including replication

[6] In my survey of students at Melbourne University, a number of subjects made their guesses about the purpose of the research explicit in the open feedback question. One example is this student, evidently puzzled by the Gettier case she considered:

> I had to think about this for a long time. Its about accepting that you can know anything. [The person in the vignette] bases their belief on previous experience [...]. But the situation that is described makes me realise that there is no way to be sure you know anything. ...

This student guessed that the experiment was about "accepting that you can know anything". The experiment's conversational context raised her standard for answering "really knows" to truly Cartesian heights: she did produce the philosophers' response, viz., that the agent only believes that p, *but for an entirely different reason to philosophers'*. This was a very common theme in the students' feedback (which is collected in the online appendix). Many who answered "only believes" claimed to have done so because, to quote another student's response, "nobody can ever truly KNOW anything".

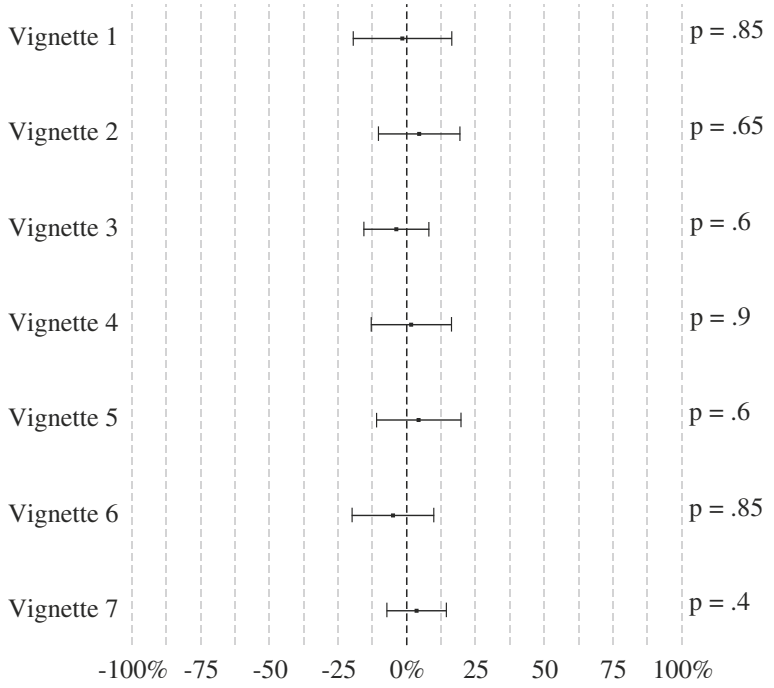

Chart 2 Estimate of percentage difference between original result and subsequent replications with 95%CIs. Note that all experiments were replicated within 5% of the original results. The vignettes referenced in this chart can be found in the online appendix

data. All p-values reported are two-tailed and calculated using Fisher's exact test; confidence intervals are 95% (exact).

3.2 Dichotomous vs. Non-dichotomous Questions

Dichotomous "yes/no" questions present subjects with mutually exclusive and exhaustive response alternatives: "Does Sarah know that p? Yes or No". Non-dichotomous questions present response alternatives which are not simply the assertion and denial of the same proposition: "Does Sarah know that p, or does she only believe it?", and "Does Sarah know that p, or does she *really* know that p?" (the latter being non-dichotomous owing to the intensifying "really").

To what extent do Weinberg et al.'s (2006) results depend on their decision to use a non-dichotomous question scheme? I hypothesised that subjects would be more likely to answer "knows" to the question "Does S know that p?", than "really knows" to the question "Does S really know that p, or does he only believe it?" For the intensifying degree adverb *really* implicates a higher standard for knowledge than simply "knows", and the downtoning adverb *only* in "only believes" implicates a lower standard for belief.

The following experiments retest a number of Weinberg et al.'s surveys, only this time their questions and response alternatives are phrased dichotomously.

In Section 3.4 we will return to consider Weinberg et al.'s arguments in light of our findings.

3.2.1 Results

I first attempted to replicate the results Weinberg et al. found when they conducted the following survey in their (2006) study:

> Bob has a friend, Jill, who has driven a Buick for many years. Bob therefore thinks that Jill drives an American car. He is not aware, however, that her Buick has recently been stolen, and he is also not aware that Jill has replaced it with a Pontiac, which is a different kind of American car.
>
> Does Bob really know that Jill drives an American car, or does he only believe it? [REALLY KNOWS/ONLY BELIEVES]

74% of Weinberg et al.'s subjects (N = 66) responded "only believes". My replication of their experiment produced effectively identical results: 71% of my subjects (N = 233) responded "only believes".

I then retested the vignette with the following difference: I asked simply "Does Bob know that Jill drives an American car?", and allowed as response alternatives just "knows" and "does not know". As expected, subjects were substantially more likely to attribute knowledge: 42% in the dichotomous group answered "knows", where only 29% had answered "really knows" in the non-dichotomous group ($p = .03$).

Intrigued by this result, I was interested to know whether the effect was robust. The experiment was repeated using Weinberg et al.'s Truetemp vignette (§2.1). Subjects in the non-dichotomous group (N = 70) were asked, following Weinberg et al., "Does Charles really know that it is 71 degrees in the room, or does he only believe it?", their response alternatives being "really knows" and "only believes". Subjects in the dichotomous group (N = 214) were asked simply "Does Charles know that it is 71 degrees in the room?", their response alternatives being "knows" and "does not know". The results were even more striking. Only 28% of subjects in the non-dichotomous group answered "really knows",[7] compared with 57% of subjects in the dichotomous group who answered "knows" ($p < .0001$)—that is, subjects were *twice as likely* to respond "knows" than "really knows" (see Chart 3).

The effect of switching from non-dichotomous to dichotomous questions and response alternatives proved robust. Further experiments returned sizeable and highly statistically significant results ($p = .0004$; $p = .0016$) (see Chart 4). I conclude that responses to Weinberg et al.'s surveys vary dramat-

[7]This result replicates Weinberg et al.'s finding: 32% of their Western subjects (N = 189) answered "really knows".

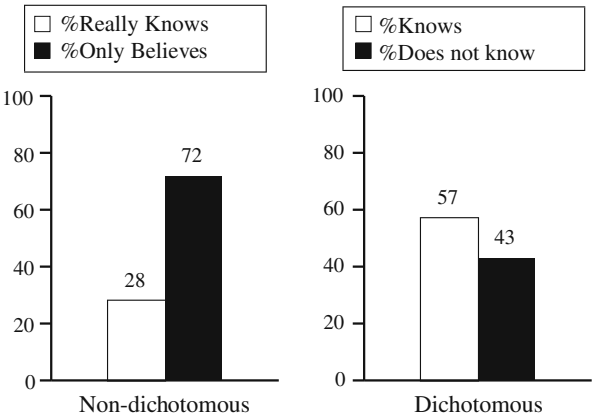

Chart 3 "Truetemp". This chart makes the effect of changing response alternatives vividly clear. The important measure, however, is the difference in the proportions of responses, i.e., 57% (knows) − 28% (really knows) = 29%. This difference is represented in Chart 4 by a point at 29% with whiskers indicating 95% confidence intervals

ically according to whether the questions and their response alternatives are phrased dichotomously or non-dichotomously.

3.3 Forced-choices and Likert Scales

In an experiment involving fifty-two undergraduates at Melbourne University, I asked subjects to comment on why they answered various forced-choice questions as they did. Their responses (reproduced in the online appendix) raised a number of subtle and occasionally rather idiosyncratic distinctions.

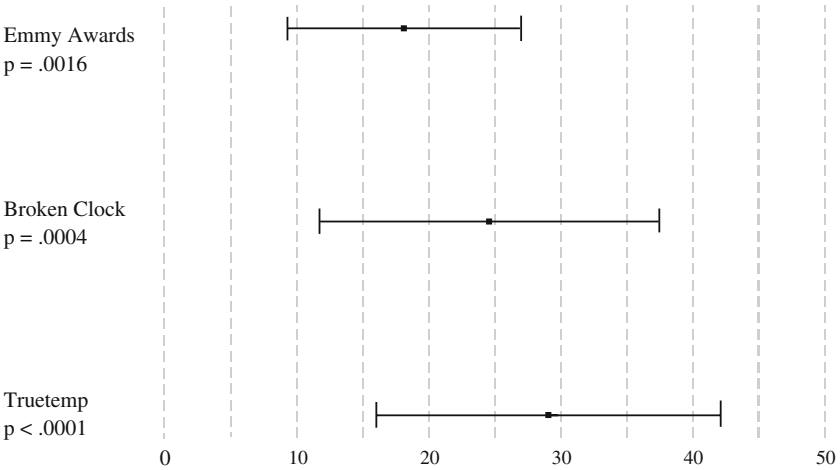

Chart 4 Estimate of percentage difference (mean 'knows' response − mean 'really knows' response) with 95%CIs. Vignettes not referenced in this text can be found in the online appendix

Consider, for example, the following students' responses to the Truetemp vignette from the previous section:

1. Charles only "believes" because he's got no reason to think he'd be spot on, it's just a guess as far as he's concerned. Perhaps if he'd been noticing on occasion after occasion that he's guessing the right temperature, then he would get to the stage where [we] could say "he knows".
2. Charles' belief is influenced by a sub-conscious activity in his brain. For his belief to become knowledge, he would have to become aware that he has some kind of crazy human thermometer power, and thus know that his brain can tell the temperature. Now it is just a strange, instinctive belief.

Lehrer (1990) meant to elicit (and of course typically succeeded in eliciting) strong intuitions *in favour* of internalism using his Truetemp thought-experiment, on which Weinberg et al.'s vignette is closely modeled. Although most subjects produced paradigm internalist type reasoning like the two examples above, there were some interesting exceptions.[8] One example is the following student:

3. The brain itself has been rewired [...] There seems to be less subjectivity with this, making it seem like the person knows it is correct, rather than it simply being opinion. If there had been other reasons why the person thought the temperature was as it was—such as they were sweating, or cold, then it would influence their belief about the actual temperature. In fact, the lack of these external factors is more likely to influence the person into thinking it is a temperature other than what they know it is.

Indeed, almost all subjects explicitly cited the internalist intuition that in order to count as knowing that p, one must be aware of one's reasons for believing p, in support of their responses. This last student, however, cited Charles' *ignorance* in support of the opposite, distinctly externalist conclusion!

Whenever experimental philosophers force subjects to respond to conceptually complex scenarios by choosing one of two mutually exclusive answers, they risk misrepresenting the structure and subtleties of folk concepts, and their subjects' understanding of the scenarios. (For an example of this last point, see Guglielmo and Malle 2008, manuscript submitted for publication). One might therefore wonder to what degree experimental philosophers' survey results depend on their choice of response format. If forced-choice results are not *roughly* replicable with Likert scales (and visa versa), we have another reason to doubt that surveys can straightforwardly reveal subjects' intuitions. Rather, they would be measuring, in part, differences in subjects' responses to another formal feature of the questionnaire.

[8]Responses were sorted by two people, one blind to the hypothesis of the study, into four categories: "internalist", "externalist", "sceptical", and "other". Agreement was high (>90%) about which responses indicated the internalist or externalist intuitions.

3.3.1 Results

In the first experiment subjects were asked:

> Two-thousand years ago, people had excellent reasons for believing that the Earth was flat. Did they know the Earth was flat? [Yes, they knew / No, they did not know].

On the left of Chart 5 are the results found with the forced-choice format, on the right are those found using a 7-point Likert scale. 86% of subjects who answered the forced-choice survey (N = 361), compared with only 34% who answered the Likert scale survey (N = 248), selected "No, they didn't know". Alternatively, counting all responses on the left-hand side of the scale (including half of those in the middle) gives 44% "Yes, they knew"—compared with only 14% in the forced-choice format. In other words, using a Likert scale increased the number of "knows" responses *three-fold* ($p < .0001$).

Further, in the forced condition over 50% of subjects reported being "very confident" of their answers and less than 15% reported being "somewhat uncertain"; in the Likert- condition only 35% of subjects (N = 248) reported being "very confident" and 26% reported being "somewhat uncertain".

The experiment was repeated using Weinberg et al.'s "Buick/Pontiac" vignette (§3.2.1) with a 5-point Likert scale. Only 32% of subjects in the Likert condition (N = 498) selected "only believes", compared with 71% of subjects who responded to the forced-choice question (N = 233) (See Chart 6).

Although it is *prima facie* plausible that Likert scales allow respondents to express their semantic intuitions more faithfully—i.e., that their concepts may not unequivocally apply or fail to apply to the case at hand—there is a plausible pragmatic explanation for these results which we should not overlook. Respondents might interpret the presence of a Likert scale as indicating that researchers regard a question as being somewhat complex. When the question appears at first blush exceedingly obvious—as my question about what people

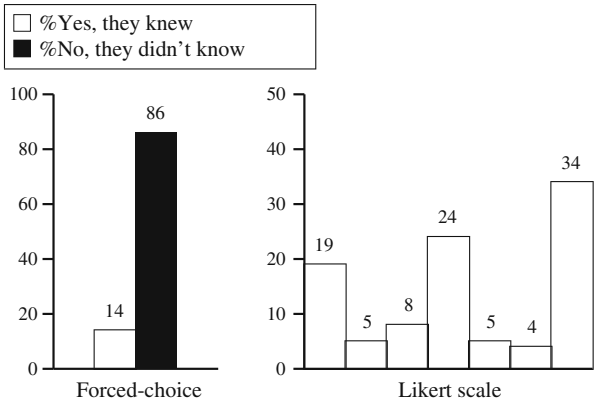

Chart 5 "Factivity". $N_{forced} = 361$; $N_{Likert} = 248$. The scale was numerically labelled with extreme anchors 'Yes, they knew' and 'No, they didn't know'

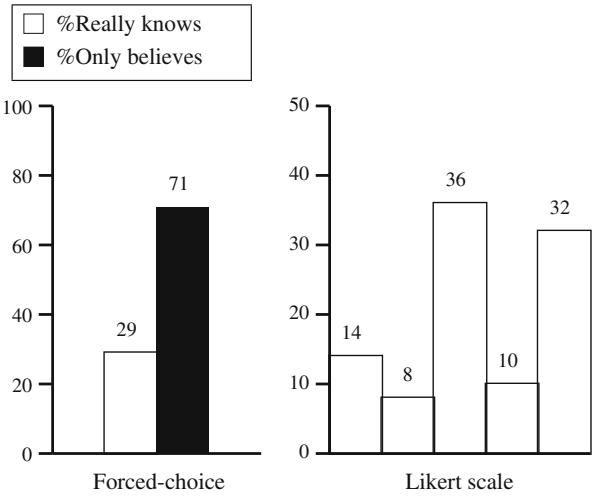

Chart 6 "Buick/Pontiac". $N_{forced} = 233$; $N_{Likert} = 498$. Numerically labelled scale with extreme anchors 'Really knows' (*left*) and 'Only believes' (*right*)

2000 years ago knew about the spherical shape of the earth must have to many respondents—this might prompt subjects to search for unintended subtleties. For as we have clearly seen, respondents draw on the formal features of a survey to help determine its intended meaning; and when asked a seemingly obvious question, people look for an alternative interpretation, one to which they can provide an intelligent response.

These experiments do not show that forced-choice results are somehow flawed. Both formats have their strengths and problems. My point is only that, whichever experimental philosophers choose, their results will reflect the decision.

3.4 Is it a "Fact" that Epistemic Intuitions Vary from Culture to Culture?

Now that we have seen how sensitive responses are to the survey's answer format, let us return to Weinberg et al.'s (2006) study where they hypothesised that "epistemic intuitions vary from culture to culture [and] from one socio-economic group to another" (ibid., p. 201). To test this hypothesis Weinberg et al. surveyed Western, East Asian, and Indian sub-continental subjects, all recruited from Rutgers University, and high and low socio-economic [SES] groups, recruited from "outside various commercial venues in down town New Brunswick …" (ibid.). After presenting subjects with a vignette describing a possible case of knowledge, all of their surveys concluded with a question of the form "Does S really know that p, or does s/he only believe it?"

Weinberg et al. found statistically significant (though frequently not terribly large) differences between the responses of Western, East Asian, and Indian sub-continental subjects, and between the responses of low- and high-SES

subjects, using a number of different survey vignettes. They conclude that their results,

> look to be yet another serious embarrassment for the advocates of IDR [Intuition Driven Romanticism] … [T]hey must either argue that these intuitive differences [between cultural groups and between SES groups] would not lead to different normative conclusions, or they must bite the bullet and argue that diverging normative claims are genuinely normative, and thus that the sorts of doxastic states that ought to be pursued by relatively rich and well educated people are significantly different from the sorts of doxastic states that poor and less well educated folks should seek. We don't pretend to have an argument showing that neither of these options is defensible. But we certainly don't envy the predicament of the IDR advocate who has to opt for one or the other (ibid., p. 212).

Yet these are clearly not the only alternatives the Intuition Driven Romanticist might opt for, since both presuppose that Weinberg et al.'s non-dichotomous forced-choice surveys were measuring subjects' epistemic intuitions. Given how sensitive their results are to the survey's formal design, an obvious alternative is that Weinberg et al. were measuring, *inter alia*, different cultures' and different SES groups' reactions to the pragmatic features of their surveys, and not just reporting "brute facts" about intuitions (ibid., p. 215).

Now perhaps Weinberg et al. might reply:

> The subjects in our inter-cultural experiments all respond to identical surveys, under identical circumstances. Therefore, when a substantial majority of East Asian subjects answer "really knows" to (e.g.) our Gettier-scenario, while the vast majority of Western subjects respond "only believes", we can be confident that it's because of their differing epistemic intuitions and not because of any pragmatic features of our surveys. It *couldn't* be due to pragmatic influences, since there is no identifiable difference in their pragmatic situations. Therefore, the Westerners and East Asians in our subject population at Rutgers University indeed have systematically varying epistemic intuitions.

Such a reply, however, would rest on a mistake. For it assumes that their Western and East Asian subjects, and low- and high-SES subjects, all respond alike to the potent pragmatic features of Weinberg et al.'s surveys. There are many reasons to believe that this assumption is false and that different groups have systematically different responses to the demands of an experimental survey. Here, then, are just three relevant examples from the current literature:

1. *Attention to pragmatic cues contained in the survey varies between collectivist and individualist cultures*. Chinese subjects "are more sensitive to conversational norms and go to more effort to observe the maxims of conversational conduct" than Western subjects (Haberstroh et al. 2002, p. 325, also see Oyserman et al. 2002). Haberstroh et al. note that ignoring these differences risks "misinterpreting cultural differences in the

question-answering process as substantive differences in the phenomenon under study" (p. 328; also see Welkenhuysen-Gybels et al. 2003 for an interesting discussion).

2. *Survey response styles vary across cultures.* Differences in cultural response styles are well-known and have been studied for over 50 years. Johnson et al. (2005, p. 267) note that since people from individualist (e.g., American) cultures tend to be "less concerned with the consequences of expressing strong opinions", they are more prone to extreme response styles (e.g., using the end-points of a Likert scale) than people from collectivist (e.g., East Asian) cultures. Conversely, response acquiescence ('yes'-saying) is lower among people from individualist cultures. Cultural differences like these are summarised by Hofstede (2001, p. 218) who notes that "respondents in a more collectivist culture are more sensitive to the social pressure they perceive to be emanating from the questionnaire".
3. *Question comprehension and literacy vary systematically between high- and low-SES groups.* Holbrook et al. (2007, p. 332) note that "years of education among adults is very strongly correlated with scores on direct tests of cognitive skills" (cf. Weinberg et al. 2006, p. 210). Further, optimally answering surveys is cognitively demanding (Krosnick 1999); especially, one imagines, in the case of experimental philosophy surveys. Since only a small number of subjects in Weinberg et al.'s study were classed as low-SES, a failure to grasp (or correctly read) their needlessly complex survey vignettes in even a moderate proportion of subjects could dramatically affect their findings.

There are many more, but these examples suffice to illustrate two things. First, we cannot assume that the differences in Weinberg et al.'s Western and East Asian, or low- and high-SES subjects' survey responses are due to underlying differences in epistemic intuitions; these differences could be due to a great many things which their methodology says absolutely nothing about. Second, at no point do Weinberg et al. consider any possibilities like those considered here; rather, they simply conclude that these groups have different epistemic intuitions. This may be the case, but from their data alone it is far from obvious.

4 Conclusion

Research has repeatedly shown that subjects rely on pragmatic cues and conversational norms to generate intelligent responses to survey questionnaires. It is only by effectively identifying intuitions with survey responses that experimental philosophers have been able to conclude that "intuitions … vary according to whether, and which, other thought experiments are considered first", or that it is a "fact that epistemic intuitions vary systematically with culture and [socioeconomic status]". These assertions are *not* made on the basis of "straightforward data about people's intuitions concerning specific

cases"; rather, they are made on the basis of how people are inclined to use certain English words like "really knows" and "only believes" within the unusual conversational context of an experimental philosophy survey.

Experimental philosophers are right to emphasise that appeals to intuition are typically empirical claims about the linguistic behaviour of competent speakers. And I think they are right to suspect that what philosophers find intuitive might diverge from what lay-people find intuitive. My concern here has not been with the motivation for experimental philosophy; rather, I want to urge experimental philosophers to re-examine their underlying methodological assumptions.

Acknowledgements For their discussion and criticism, I am grateful to Joshua Alexander, Sue Finch, Frank Jackson, Tania Lombrozo, Edouard Machery, Thomas Nadelhoffer, Norbert Schwarz, Gerard Vong, Jonathan Weinberg, and especially to Neil Thomason and Ilana Payes.

References

Alexander, Joshua and Jonathan M. Weinberg. 2007. Analytic epistemology and experimental philosophy. *Philosophy Compass* 2(1):56–80.

Ceci, Stephen J. 1991. How much does schooling influence general intelligence and its cognitive components?: A reassessment of the evidence. *Developmental Psychology* 27:703–722.

Doris, John M., Joshua Knobe, and Robert L. Woolfolk. 2007. Variantism about responsibility. *Philosophical Perspectives* 21(1):183–214.

Goldman, Alvin and Joel Pust. 1998. Philosophical theory and intuitional evidence. In *Rethinking intuition*, eds. William Ramsey and Michael DePau. Totowa: Rowman & Littlefield.

Haberstroh, Susanne, Daphna Oyserman, Norbert Schwarz, Ulrich Kuhnen, and Li Jun Ji. 2002. Is the interdependent self more sensitive to question context than the independent self? Self-construal and the observation of conversational norms. *Journal of Experimental Social Psychology* 38:323–329.

Higgins, Edward Tory and Arie Kruglanski, eds. 1996. *Social psychology: Handbook of basic principles*. Newyord: Guilford.

Hofstede, Geert H. 2001. *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. London: Sage.

Holbrook, Allyson L., Jon A. Krosnick, David Moore, and Roger Tourangeau. 2007. Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly* 71(3):325–348.

Jobe, Jared B. and David J. Mingay. 1991. Cognition and survey measurement: History and review. *Applied Cognitive Psychology* 5:175–192.

Johnson, Timothy, Patrick Kulesa, Isr Llc, Young Ik Cho, and Sharon Shavitt. 2005. The relation between culture and response styles. *Journal of Cross-Cultural Psychology* 36:264–277.

Kauppinen, Antti. 2007. The rise and fall of experimental philosophy. *Philosophical Explorations* 10(2):95–118.

Knobe, Joshua. 2006. The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies* 130:203–231.

Knobe, Joshua. 2007. Experimental philosophy. *Philosophy Compass* 2:81–92.

Krauss, Robert M. and Susan R. Fussell. 1996. Social psychological models of interpersonal communication. In *Social psychology: Handbook of basic principles*, eds. Edward Tory Higgins and Arie Kruglanski. Newyord: Guilford.

Krosnick, Jon A. 1999. Survey research. *Annual Review of Psychology* 67:537–567.

Krosnick, Jon A. and Michael A. Milburn. 1990. Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgements. *Journal of Social Psychology* 54:940–952.

Lehrer, Keith. 1990. *Theory of knowledge*. Boulder: Westview.

Machery, Edourd, Ron Mallon, Shaun Nichols, and Stephen P. Stich. 2004. Semantics, cross-cultural style. *Cognition* 92(3):B1–B12.

Nadelhoffer, Thomas and Eddy Nahmias. 2007. The past and future of experimental philosophy. *Philosophical Explorations* 10(2):123–149.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and J. Turner. 2005. Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18(5): 561–584.

Nichols, Shaun. 2004. Folk concepts and intuitions: from philosophy to cognitive science. *Trends in Cognitive Science* 8(11):514–518.

Nichols, Shaun, Stephen P. Stich, and Jonathan M. Weinberg. 2003. Metaskepticism: Meditations in ethno-epistemology. In *The skeptics: Contemporary essays. Ashgate epistemology and mind series*, ed. Steven Luper. Aldershot: Ashgate.

Norenzayan, Ara and Norbert Schwarz. 1999. Telling what they want to know: participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology* 29:1011–1020.

Oyserman, Daphna, Heather M. Coon, and Markus Kemmelmeier. 2002. Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin* 128(1):3–72.

Schaeffer, Nora Cate and Stanley Presser. 2003. The science of asking questions. *Annual Review of Sociology* 29:65–88.

Schwarz, Norbert. 1995. What respondents learn from questionnaires: The survey interview and the logic of conversation. *International Statistical Review/Revue Internationale de Statistique* 63(2):153–168.

Schwarz, Norbert. 1996. *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Mahwah: Erlbaum.

Schwarz, Norbert, Fritz Strack, and Hans-Peter Mai. 1991. Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *The Public Opinion Quarterly* 55:3–23.

Strack, Fritz, Leonard L. Martin, and Norbert Schwarz. 1988. Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology* 18:429–442.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1995. *Thinking about answers: The application of cognitive process to survey methodology*. Jossey-Bass.

Swain, Stacey, Joshua Alexander, and Jonathan M. Weinberg. 2008. The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and Phenomenological Research* 76(1):138–155.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge: Cambridge University Press.

Weinberg, Jonathan M., Shaun Nichols, and Stephan Stich. 2006. Normativity and epistemic intuitions. In *Biological and cultural bases of human inference*, eds. Riccardo Viale, Daniel Andler, and Lawrence A. Hirschfeld. Mahwah: Lawrence Erlbaum.

Welkenhuysen-Gybels, Jerry, Jaak Billiet, and Bart Cambré. 2003. Adjustment for acquiescence in the assessment of the construct equivalence of likert-type score items. *Journal of Cross-Cultural Psychology* 34(6):702–722.